DOCUMENT RESUME

ED 409 377 TM 026 957

AUTHOR Lunz, Mary E.
TITLE Performance Examinations: Technology for Analysis and Standard Setting.
PUB DATE Mar 97
NOTE 24p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Computer Assisted Testing; *Criteria; *Educational Technology; *Estimation (Mathematics); *Performance Based Assessment; Scaling
IDENTIFIERS *FACETS Computer Program; Fair Average Method; *Standard Setting

ABSTRACT

This paper explains the multifacet technology for analyzing performance examinations and the fair average method of setting criterion standards. The multidimensional nature of performance examinations requires that multiple and often different facets elements of a candidate's examination form be accounted for in the analysis. After this is accomplished, all candidate ability estimates are located on the same scale. The fair average standard setting method, while substantially different from more traditional methods, provides a criterion standard in a score metric that is easy for most people to understand. Yet, it accounts for the influence of the particular facets elements in each test form, and can be used to establish a pass point that applies appropriately to all candidates regardless of the raters or tasks or problems on each candidate's examination. Performance examinations are extremely complex because of the number of possible examination forms that may occur during an administration. The technology provided by computer programs, as exemplified by the FACETS program, is essential to sorting out the impact of each facet element on candidate scores and setting a criterion standard that accounts for the impact of all examination facets. (Contains two figures, three tables, and nine references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 409 377

PERFORMANCE EXAMINATIONS:

TECHNOLOGY FOR ANALYSIS AND STANDARD SETTING

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Mary Lunz

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Mary E. Lunz, Ph.D.

American Society of Clinical Pathologists

TM026957

Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, Illinois, March, 1997.

BEST COPY AVAILABLE

PERFORMANCE EXAMINATIONS:

TECHNOLOGY FOR ANALYSIS AND STANDARD SETTING

Abstract

The purpose of this paper is to explain the multi-facet technology for analyzing performance examinations and the fair average method of setting criterion standards. The multi-dimensional nature of performance examinations requires that multiple and often different facet elements of a candidate's examination form be accounted for in the analysis. After this is accomplished, all candidate ability estimates are located on the same scale. The fair average standard setting method, while substantially different than more traditional methods, provides a criterion standard in a score metric that is easy for most people to understand. Yet, it accounts for the influence of the particular facets elements in each test form, and can be used to establish a pass point that applies appropriately to all candidates regardless of the raters or tasks or problems on each candidate's examination.

PERFORMANCE EXAMINATIONS:

TECHNOLOGY FOR ANALYSIS AND STANDARD SETTING

The purpose of a performance examination is to infer candidate abilities beyond the particular sample of tasks, projects and raters, etc. on the examination. The candidate ability is then compared to a standard, and pass/fail decisions are made. The performance examination should be designed to measure candidate ability as accurately and consistently as possible, so that all successful candidates must demonstrate an established level of knowledge and skill. Performance examinations, historically, have not yielded reproducible pass/fail decisions. Two related issues contribute to this situation, namely, availability of : 1) technology to control for the rater specific or other contextual bias and 2) standard setting methods that can be used on multi-facet performance examinations.

A better understanding of the structure of a performance examination, is achieved by breaking the examination into its component parts or facets, so the influence of each facet on the score can be observed. The basis for validity of any examination is the meaning assigned to the scores (Messick, 1995); therefore, it is helpful to understand, as fully as possible, how the score of a performance examination is derived from the combination of facets elements in an examination form. For example, one candidate may be rated by severe raters on difficult projects, while another candidate may be rated by more lenient raters on projects of moderate difficulty. This is the contextual bias of performance examinations. These differences must be accounted for in the scoring system and in the standard setting process to insure consistent measurement among candidates and valid inferences from the scores.

Facets of a Performance Examination

Performance examinations necessitate accounting for at least three facets. Usually, ratings are from scales that have more than two points, and there are always more than the two facets (items and candidates).

There are typically four separate facets or components that make-up a performance examination. The first facet is candidate ability, which encompasses the knowledge and skill possessed by the candidate with regard to the problem, task or product measured by the performance examination. It is expected that candidates will vary in their ability.

The second facet is the cases or projects. Some projects have detailed specifications that are comparable across candidates performances. Examples are medical cases, essay prompts, science, or laboratory projects. The requirements are described to candidates who then perform to the best of their ability. Other performance examinations, allow candidates to select a sample of their own work. In medicine, candidates may select cases from their medical practice to present in an oral examination. Portfolios may be developed in art or writing. The performances usually cover specific content specifications so that general areas of knowledge and skill are represented. Sometimes cases are structured and all candidates are challenged by the same group of cases.

The third facet is the severity of the rater. Raters are essential for performance examinations; however, raters have unique physical and mental characteristics, as well as, unique reactions to the examination, all of which influence their ratings. Raters tend to have internalized standards, so some raters give consistently lower ratings across candidates while others tend to be more generous (Lunz and O'Neill, 1997). Training focuses and directs a rater's attention, but is

usually unable to alter permanently the knowledge and skill that has developed over a lifetime (Lunz and Stahl, 1993).

The fourth facet is the tasks associated with each project or case. Considerations for this facet include: (1) the number of tasks rated, (2) the extent of detail in the definition of the tasks, and (3) the relevance of the tasks to the cases or projects. Tasks may be fairly objective, such as using correct punctuation, or may be more subjective such as ethical standards for medical treatment. Tasks must be carefully delineated.

The fifth component is the definition of the rating scale. Rating scales provide a "disciplined dialogue" which encourages raters to assign specified meaning to each category on the rating scale. Rating scales may have as few as two categories (0/1) or an infinite number (0/∞). Usually, each category on the scale has a specific definition. The definitions of the rating scale categories are important, because they influence how raters assign points to the quality of the performance. The measurement distance among the rating scale categories impacts the ratings given to candidates. For example, there is a great distance between "unacceptable" and "excellent," so logical categories between these extremes are often inserted, (e.g. marginal, acceptable). The categories of the rating scale provide focus for the rater. The overall rating commensurate with "satisfactory" or "competent" to practice most likely represents the region where a criterion standard and pass point will be established. The standard setting process changes the value judgment of satisfactory into a score that is believed by experts to distinguish between candidates who have sufficient knowledge and skill to pass and those who do not.

Subjectivity and Performance Examinations

Performance examinations depend upon the ratings awarded by raters. Hopefully, the raters have been trained in the use of the rating scale, and have appropriate qualifications with regard to knowledge, experience, understanding and judgment. The more assessments made by the raters, the closer the score is likely to be to "true" ability. As more ratings are accumulated, the precision of the measurement increases, and the error of measurement associated with a candidate's score decreases. Low measurement error increases confidence in the pass/fail decisions.

The examination design determines the number of ratings per candidate. One holistic rating yields a high measurement error. The rater must make many preliminary mental decisions to arrive at the final judgment, and each rater probably uses a different strategy. It is possible that most raters have less than absolute confidence in their rating when all aspects of candidate performance is condensed into one rating. Multiple ratings awarded on tasks within cases or projects encourage the rater to be more analytic when assessing the candidate's performance. This also reduces the error of measurement, and increases confidence in the accuracy of the final score.

TECHNOLOGY FOR ANALYZING PERFORMANCE EXAMINATIONS: IRT MODELING

Background and Explanation of the Technology of the Multi-Facet Model

Item response theory modelling has been used successfully for examination analysis and equating. The basic Rasch Model (Rasch, 1960/1980) is a mathematical representation of the candidate and item interaction. The probability of a candidate answering a particular item

correctly is modeled as:

$$\log(P_{ni}/(1 - P_{ni})) = (B_n - D_i) \quad (1)$$

where: (P_{ni}) is the probability of answering the item correctly

$(1 - P_{ni})$ is the probability of answering the item incorrectly

B_n is the ability of the candidate n

D_i is the difficulty of the item i

The probability of a correct response is a function of the difference between the ability of a candidate and the difficulty of the item. If a candidate's ability is greater than the difficulty of the item, then the probability of answering correctly is greater than 50%. If the difficulty of the item is greater than the ability of the candidate, the probability of answering the item correctly is less than 50%. The use of the logarithmic function in the equation transforms ordinal ratings to a linear scale. The unit of measure is log-odds units or "logits" (Wright and Stone, 1979).

For analysis of performance examinations the basic Rasch model is extended to the multi-facet model (Linacre, 1989), so that facets for task and item difficulty as well as rater severity, can be added to the equation. Severity is the term used to encompass the factors that influence the way raters rate candidate performances. Thus, when a candidate is rated, the probability of a candidate succeeding is modelled:

$$\log((P_{nmjik}) / P_{nmjik-1})) = (B_n - T_m - C_j - D_i - R_k) \quad (2)$$

where: (P_{nmjik}) is the probability of being rated in category k

$(P_{nmjik-1})$ is the probability of being rated in category k-1

B_n is the ability of the candidate n

T_m is the difficulty of the task m

C_j is the severity of the rater j

D_i is the difficulty of the project i

R_k is the difficulty of being rated in category k rather than category k-1

The probability of receiving a rating is a function of the difference between candidate ability and the task difficulty, after adjustment for the severity of the rater(s) and the difficulty of the cases/projects. If the candidate's ability is higher than the difficulty of the projects, after adjustment for the task difficulty and the rater severity, then the probability of a satisfactory rating is greater than 50%. Conversely, if the task difficulty after adjustment for rater severity, is greater than the ability of the candidate, the probability of receiving a satisfactory rating is less than 50%.

The ordering of the candidates, tasks, raters, and projects on a linear scale provides a frame of reference for understanding the relationship of the facets of the performance examination. It makes it possible to observe estimated candidate ability (B_n) from highest to lowest, estimated task difficulty (T_m) from most to least difficult, estimated rater severity (C_j) from most to least severe, and estimated project difficulty (D_i) from most to least difficult (see Figure 2).

The ratings are the basic unit of analysis. Task difficulty is calculated from all ratings given to all candidates by all raters on a task. Project difficulty includes all ratings associated with specific projects given by all raters across all candidates. Rater severity is calculated from all the ratings given by a specific rater. Thus, all estimates for ability, difficulty and severity are derived from sums of ratings, as shown in Figure 1, so a linking system must be included in the examination design to insure an accurate representation of the relationship of the facets of the examination..

The multi-facet model converts raw score ratings into log odds units or logits, thus creating an equal interval scale. Candidate ability estimates, rater severity estimates, task and/or project difficulty estimates are reported in logits, and each facet has a mean of zero. Estimates can range from 0 to ± ∞, usually with the scale set up so that **high positive** indicates more severe, more able, more difficult and **high negative** means lenient, less able or easy.

The multi-facet model also provides estimates of the consistency of the ratings for candidates, raters, projects, and tasks. This is reported as the fit of the data to the model. The model expects observed ratings to be consistent, that is, able candidates should earn higher ratings than less able candidates, from all raters on most/all tasks within projects. More difficult projects or tasks should cause lower ratings to be awarded than easier projects or tasks by all raters. Fit statistics indicate inconsistent ratings on any of the facets. For example, the fit statistics for raters indicate the degree to which each rater is internally self-consistent across candidates, tasks, and projects (intra-rater consistency). Raters who award unexpectedly high or low ratings to a particular candidate on a particular task or project are identified and the effect of the unexpected ratings on the candidate ability estimates reviewed. The fit statistic for each project and task indicates inter-rater consistency, or perhaps an examinee who does something totally unexpected. Misfit indicates that some raters deviated significantly from others when grading the project or task for some candidates. Criteria for acceptable fit are situationally dependent. Mean-squared residual differences between .5 and 1.5 are often criteria for acceptability.

FACETS: the computer program that provides the technology

FACETS (Linacre, 1996) a computer program designed to analyze data from performance examinations uses the principles of the multi-facet model to calibrate the elements of each facet and then to account for differences in the examinations of each candidate before their ability is estimated. The FACETS program can yield reasonable calibrations of facet elements only when there is a reasonable level of standardization within the examination design and reasonable overlap or linking among examination facets. The **rating scale** must be standardized. All raters must use the same rating scale categories, understand the definitions of the rating scale categories, and be willing to use the categories to the best of their ability when rating the quality of candidate performance. The rating scale is the foundation for scoring candidates.

Also the **tasks or skills** rated should be comparable across projects or portfolios and be defined explicitly. If three tasks are defined, then all raters must be willing to assess candidates on the three tasks. If six tasks are defined, then all raters must use the six tasks to rate candidate performance, unless otherwise defined. Other facets may have differing levels of standardization depending upon the design of the examination. When the requirements are met, the FACETS program provides the technology for performance examination analysis and the opportunity to establish criterion standards that must be achieved by all candidates who pass the examination. Figure 1 shows typical overlap patterns among facets. Note that tasks and projects overlap completely in this example, while candidates and raters do not overlap completely causing different forms of the examination to be constructed. Of course, the more facets included in the examination, the more complex the linking patterns.

INSERT FIGURE 1 ABOUT HERE

STANDARD SETTING USING MULTI- FACET TECHNOLOGY

All item response theory models sum the ratings to create total scores that are meant to represent the level of candidate performance. However, on a performance examination, raw scores do not take into account differences in the difficulty of the examination forms each candidate challenged. Examination forms differ among candidates because their performance is graded by different raters, perhaps on different problems or prompts. The multi-facet technology calculates and includes a correction factor that accounts for the particular characteristics of the examination form encountered by the candidate including project and task difficulty, as well as, rater severity, before candidate ability estimates are calculated. A scale is constructed on which all candidates are positioned. The examination forms have, essentially, been equated, because differences in examination forms have been accounted for in the analysis. Once the scale is established, a criterion standard can be established on the scale. Candidates who meet the standard pass, others fail, but all candidates must demonstrate the same level of ability to pass.

A criterion standard usually answers the question "how much is enough?" That is, how much knowledge and skill should be required to distinguish those who should pass from those who should fail. Usually, the rating scale associates a rating with satisfactory performance (e.g. satisfactory = 3) and each rater applies that rating scale knowing that "3" means performance is satisfactory to pass. However, even well trained raters often have somewhat different concepts of satisfactory performance, so raw score averages provide an advantage for candidates who get lenient raters.

Candidate ability estimates presented in log-odds units account for test form differences, but are often difficult to understand, while average scores are commonly understood. For this reason, the FACETS program provides the FAIR AVERAGE calculation.

$$\text{FAIR AVERAGE} = \text{RSR} * (e^{\text{LPC}} / 1 - e^{\text{LPC}}) \quad (3)$$

where:

RSR = rating scale range (n of categories)
LPC = logit percent correct
B_o = observed ability
B_n = ability estimate corrected for contextual bias
m = rating scale centering [depends on n of categories]
x = rating scale slope [intercept]

$$B_o = m + x * \log (R/W)$$

$$B_n = B_o - D_i - C_j - T_m \quad \text{(see formula 2)}$$

$$LPC = (B_n - m) / x$$

The FAIR AVERAGE is the candidate's transformed score after correction for the contextual bias of the examination form and the addition of a scaling factor. This removes the effect of the contextual bias from a **satisfactory** rating, so a fair average of 3.00 or **satisfactory** means that the candidate demonstrated satisfactory performance regardless of the difficulty of the examination form taken. What makes the average **fair** is that it is <u>not</u> biased by the rater severity or the difficulty of the tasks or problems, so satisfactory, has the same interpretation across all candidates. The fair average is, directly linked to the meaning of the rating scale category and can therefore be used as an absolute criterion standard. **The question of how much is enough is answered by a fair average score that represents satisfactory knowledge and skill to practice in the field.**

INSERT FIGURE 2 ABOUT HERE

DEMONSTRATION ANALYSIS

The data used in this demonstration analysis are from a medically related interview (oral) examination. The data set had four facets: 1) candidates; 2) raters; 3) topics; and 4) tasks.

Three tasks were rated: 1) Recall; 2) Interpretation (Interp); and 3) Problem solving (PS). A five point rating scale was used in which 4 = excellent, 3 = good, 2 = satisfactory, 1 = marginal and 0 = unsatisfactory. Three pairs of two raters met with and rated a candidate's performance, independently , for a total of six raters per candidate. The first pair of raters assessed the candidate on Topic 1, Biology (Bio), the second pair of raters assessed the candidate on Topic 2, Chemistry (CHEM), and the third pair of raters assessed on Topic 3, Physiology (PHY). The raters rotated so that all raters rated all topics during the course of the examination. The pairs of raters also rotated so candidates encountered different pairs of raters for each topic. All raters rated candidates on the same tasks within topics. This created the overlap necessary for the FACETS analysis because candidates had all topics, all tasks and some raters in common, and raters had all topics and tasks and some candidates in common. The same rating scale was used by all raters for all tasks, for all candidates.

Data were analyzed using the FACETS computer program (Linacre, 1996) first to describe the characteristics of the examination and then to assist in setting a criterion standard. Figure 2 shows the relative position of the elements within each examination facet. Table 1 shows the raters in severity estimate order. Table 2 shows the topics and tasks in difficulty estimate order. Since all candidates challenged all topics and tasks, the adjustment is dictated primarily by the severity of the rater. Table 3 shows the equated candidate ability estimates. Ability estimates ranged from 4.98 to -1.30 logits after correction for raters. Raters agreed on the quality of candidate performance for all but eight candidates (see fit statistics *). The candidates on which there was disagreement are in all quartiles of the ability distribution. Most of the ratings (52%) indicated candidate performance to be "good" (see Table 2). Overall, the raters were able to distinguish among candidate abilities. The

candidate separation reliability (comparable to Cronbach's Alpha) is .92. This indicates that the ability estimates reliably differentiate among candidates.

A Fair Average of 2.0 (Satisfactory) translates to a logit ability estimate of approximately -.32 on the candidate ability scale (see Figure 2). This can be adjusted by the measurement error to avoid passing a candidate who should fail or failing a candidate who should pass. The standard setting committee must make the final decision about the placement of the pass point.

In this example only 5% of the ratings given were less than 2 (Satisfactory). The Fair Average standard setting method indicates that a high percentage of the candidate sample can meet the criterion standard. This could indicate a very able candidate sample, lenient raters or limited use of the rating scale. If the standard is adjusted by the error of measurement, the confidence in the accuracy of the decision is determined. In this example, it may be desirable to have 95% confidence that no candidate who should fail would pass. This moves the pass point to .34 logits (1.65 (.40) = .66 + -.32 = .34). The statistics describe the region of the standard, the standard setting committee identifies the actual pass point on the scale.

INSERT TABLES 1, 2, 3 AND FIGURE 2 ABOUT

Discussion and Comments

The use of the multi facet model and the FACETS computer program shows the relationships among facets, and provides a strategy for setting a reasonably objective criterion standard that is directly linked to the rating scale categories. It is important to have appropriate words that delineate the differences between passing and failing performance. Planning the examination carefully,

defining the rating scale and familiarizing raters with the examination process all contribute to the interpretation of the criterion standard. When the Fair Average is used to set the standard , a passing candidate must demonstrate satisfactory performance regardless of the raters who assess candidate performance.

Cizek (1996) provides guidelines for establishing criterion-referenced standards. While many of the issues appear to be directed toward written multiple choice examinations, the basic guidelines can be used to evaluate the fair average method of standard setting for performance examinations.

The first issue is the adequacy and applicability of the method. The fair average method produces a standard that is easy to explain, define and connect to the rating scale; however, it is readily available only when the multi-facet technology is used to analyze performance exams.

The second issue is the qualification of standard setting participants. By providing input on the performance of the candidates they grade, all raters have direct input into the standard setting process. Raters are usually selected because they have the necessary qualifications, experience and training to rate candidate performances. The more input into the standard the more likely it will be representative of satisfactory performance.

The third issue is reliability and validity of the cut score. When the fair average method is used, the validity of the standard setting process is the validity of the examination because both occur simultaneously. We tend to believe that performance exams have validity because they are direct observations of candidate performance. The standard is also based on direct observation of candidate performance. The fair average method is reliable because it is based on the multi-facet adjustment and scaling that removes the contextual bias from each examination form. The standard is independent of the particular raters that gave the rating, so it can be correctly interpreted as

meaning satisfactory.

The fourth issue is documentation of the process. Because the fair average standard setting process is interwoven within the examination administration, and there are no extra sessions or data collection meetings, the documentation for the standard setting process is essentially the same as the documentation for the exam administration and analysis. The interpretation and use of the fair average must be documented with regard to establishing the criterion standard, but no additional transformations of the data are necessary unless it is decided to adjust the fair average standard by the standard error of measurement (SEM) to insure some level of confidence in the pass/fair decisions.

The fifth issue is accessibility of the process. The easiest way to use the fair average standard setting method is to use the multi-facet model and FACETS program. This limits users of this standard setting method to owners of the FACETS program. However, the multi-facet model is an item response theory model which can be programmed, and fair average can be calculated using standard techniques (see p. 11 formula 3). It would be complex, but possible. In summary, the fair average method relates directly to the examination administration, but is somewhat limited by access to the multi-facet technology.

A limitation of using the Fair Average for standard setting is that the same rating scale must be used for all ratings on the examination. If different rating scales are used for different tasks or projects, etc. the interpretation of the standard is less clear cut because it incorporates several different scale definitions.

The advantage of using the multi-facet model is that all facets of the examination are mapped so that the relative placement of facet elements is observable. It becomes possible to define the characteristics of the particular examination taken by each candidate. Being able to identify the characteristics of a candidate's examination makes it possible to understand the rationale for the adjustment. Each examination form is traceable. In addition, an estimate of the reliability of candidate separation is included. This is an important reliability estimate (comparable to Cronbach Alpha) because it documents that the examination reliably differentiates among candidate ability levels.

The major advantage of setting a criterion standard for a performance examination is that it is established by a group of experts rather than each individual rater as he/she rates individual candidates. The definition of the standard can be published, and the standard can be applied consistently among candidates. This is extremely important for certification and other high stakes examinations.

Performance examinations are extremely complex because of the number of possible examination forms that may occur during an administration (Form = raters + topics + tasks + etc.). This makes both analysis and standard setting extremely complex. Therefore, the technology provided by computer programs is essential to sorting out the impact of each facet element on candidate scores and setting a criterion standard that accounts for the impact of all examination facets.

FIGURE 1

CALCULATION OF PATTERNS OF OVERLAP REQUIRED FOR THE FACET ELEMENT CALIBRATIONS

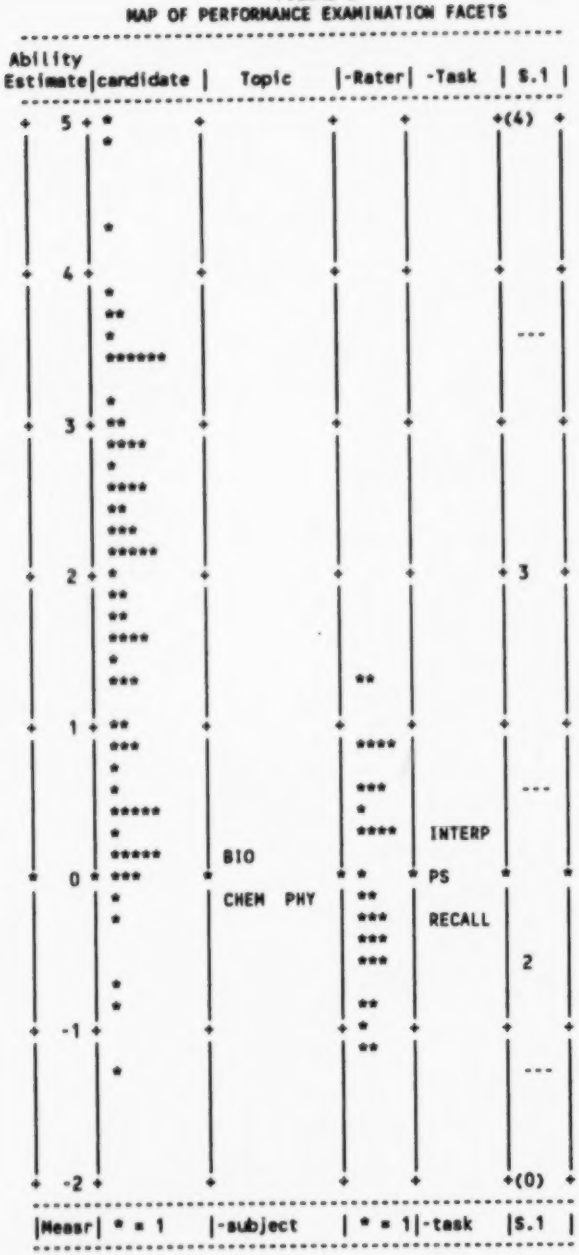
	Candidates		Tasks		Projects		Raters
C_j = Raters	**SUBSET**	+	ALL	+	ALL	=	**CALIBRATION**
D_i = Projects	ALL	+	ALL	=	**CALIBRATION**	+	ALL
T_m = Tasks	ALL	=	**CALIBRATION**	+	ALL	+	ALL
B_n = Candidates	**CALIBRATION**	=	ALL	+	ALL	+	**SUBSET**

CALIBRATION - Candidate ability, Tasks or Project difficulty, Rater severity estimate

SUBSET - raters interact with a subset of candidates and/or candidates interact with a subset of raters.

ALL - all tasks and/or projects are rated by all raters, and/or challenged by all candidates.

FIGURE 2
MAP OF PERFORMANCE EXAMINATION FACETS



Shows the relationship of the facets of this performance examination
* Represents a candidate ability or rater severity estimate

BEST COPY AVAILABLE

TABLE 1

RATERS IN SEVERITY ORDER

Rater Number	Obsvd Score	Num of Rating	Observed Average	Fair Avrage	Rater Severity	S.E.	Infit MnSq	Outfit MnSq
					Most Severe			
34	116	42	2.8	1.5	1.36	.25	1.1	1.1
3	77	33	2.3	1.5	1.33	.26	0.9	1.0
27	113	45	2.5	1.8	.90	.23	1.4	1.3
29	109	42	2.6	1.8	.90	.24	0.4	0.4
18	109	45	2.4	1.8	.89	.22	0.6	0.7
9	125	45	2.8	1.8	.85	.24	0.7	0.8
8	118	45	2.6	2.0	.59	.23	1.2	1.2
11	114	45	2.5	2.0	.59	.23	1.4	1.5
16	117	42	2.8	2.0	.51	.25	0.8	0.9
25	123	48	2.6	2.0	.48	.22	1.0	0.9
1	119	42	2.8	2.1	.27	.26	0.8	0.8
12	130	45	2.9	2.1	.27	.25	1.4	1.4
15	127	45	2.8	2.2	.25	.24	0.6	0.6
10	127	45	2.8	2.2	.24	.25	1.3	1.3
20	130	45	2.9	2.3	.00	.25	1.0	1.0
33	127	42	3.0	2.3	-.16	.27	0.6	0.6
28	122	42	2.9	2.4	-.17	.26	1.5	1.6
21	129	45	2.9	2.4	-.23	.25	0.7	0.7
23	137	45	3.0	2.4	-.29	.26	0.5	0.5
13	139	45	3.1	2.4	-.35	.26	1.4	1.3
26	142	45	3.2	2.5	-.44	.27	0.5	0.5
32	135	45	3.0	2.5	-.47	.26	0.9	1.1
7	138	45	3.1	2.5	-.50	.26	1.2	1.2
30	128	42	3.0	2.5	-.56	.27	0.7	0.8
31	139	45	3.1	2.5	-.58	.26	1.1	1.1
6	145	48	3.0	2.5	-.62	.25	1.4	1.3
24	137	45	3.0	2.6	-.81	.26	0.9	0.9
22	138	42	3.3	2.6	-.90	.29	0.9	1.0
19	85	27	3.1	2.7	-.94	.34	1.0	1.0
14	142	45	3.2	2.8	-1.20	.27	1.0	1.0
4	105	30	3.5	2.8	-1.21	.39	1.9	2.2*
					Least Severe			
Mean	123.9	43.0	2.9	2.2	.00	.26	1.0	1.0
S.D	15.5	4.6	0.3	0.3	.71	.03	0.4	0.4

RMSE .26 Adj S.D. .66 Separation 2.51 Reliability .86
Fixed (all same) chi-square: 223.9 d.f.: 30 significance: .00

Rater Severity presented in logits.

*Only Rater #4 was inconsistent. Since he was very lenient overall, it is likely that a lower than expected rating was given to a candidate.

TABLE 2
TASKS AND TOPICS IN DIFFICULTY ORDER

TASKS IN DIFFICULTY ORDER

N	task	Obsvd Score	Num of Ratings	Average Score	Fair Avrge	Diff Estimate	Model S.E.	Infit MnSq	Outfit MnSq
2	INTERPRETATION	1240	444	2.8	2.1	.25	.08	1.0	1.0
3	PROBLEM SOLVING	1271	444	2.9	2.2	.07	.08	0.9	1.0
1	RECALL	1331	444	3.0	2.4	-.32	.08	1.0	1.0
Mean (Count: 3)		1280.7	444.0	2.9	2.3	.00	.08	1.0	1.0
S.D.		37.8	0.0	0.1	0.1	.24	.00	0.0	0.0

RMSE .08 Adj S.D. .23 Separation 2.84 Reliability .89
Fixed (all same) chi-square: 26.8 d.f.: 2 significance: .00

TOPICS IN DIFFICULTY ORDER

N	Subject	Obsvd Score	Num of Ratings	Average Score	Fair Avrge	Diff Estimate	Model S.E.	Infit MnSq	Outfit MnSq
2	BIOLOGY	1251	444	2.8	2.2	.19	.08	1.0	1.0
1	CHEMISTRY	1295	444	2.9	2.3	-.09	.08	1.1	1.1
3	PHYSIOLOGY	1296	444	2.9	2.3	-.10	.08	0.9	0.9
Mean (Count: 3)		1280.7	444.0	2.9	2.3	.00	.08	1.0	1.0
S.D.		21.0	0.0	0.0	0.1	.13	.00	0.1	0.1

RMSE .08 Adm S.D. .11 Separation 1.34 Reliability .64
Fixed (all same) chi-square: 8.6 d.f.: 2 significance: .01

USE OF RATING SCALE CATEGORIES*

Response Category Name	DATA Score	Used	%	Cum. %	STEP CALIBRATIONS Measure	S.E.	Meaning
0 SATISFACTORY	0	14	1%	1%			
1 MARGINAL	1	50	4%	5%	-2.09	.29	Very Easy
2 SATISFACTORY	2	297	22%	27%	-1.77	.15	Easy
3 GOOD	3	686	52%	79%	.38	.08	More Difficult
4 EXCELLENT	4	285	21%	100%	3.48	.08	Very Difficult

*Only 5% of ratings were Unsatisfactory or Marginal.

BEST COPY AVAILABLE

TABLE 3

CANDIDATES IN ABILITY ESTIMATE ORDER WITH FAIR AVERAGE STANDARDS INDICATED

Candidate	Obsvd Score	Average Score	Fair* Avrge	Ability Estimate	S.E.	Infit MnSq	Outfit MnSq	
405	67	3.7	3.8	4.98	.53	1.2	1.4	
412	69	3.8	3.8	4.82	.63	1.1	1.3	
304	63	3.5	3.7	4.22	.46	1.7	2.7*	
305	60	3.3	3.6	3.90	.43	1.8	1.8	
506	62	3.4	3.5	3.73	.45	0.4	0.4	
505	62	3.4	3.5	3.72	.45	0.9	0.9	
104	62	3.4	3.5	3.62	.46	0.6	0.5	
214	62	3.4	3.5	3.49	.45	1.3	1.3	
211	62	3.4	3.5	3.47	.45	1.0	1.0	
513	62	3.4	3.5	3.46	.44	1.5	1.4	
515	62	3.4	3.5	3.46	.44	0.9	0.8	
112	63	3.5	3.4	3.40	.45	0.8	0.8	
308	61	3.4	3.4	3.37	.43	0.9	0.9	
504	59	3.3	3.4	3.12	.43	1.0	1.1	
310	63	3.5	3.3	2.98	.44	0.9	0.9	
311	63	3.5	3.3	2.98	.44	0.8	0.8	
408	60	3.3	3.3	2.93	.43	1.7	1.6*	
208	60	3.3	3.3	2.90	.43	1.0	1.1	
512	55	3.1	3.3	2.87	.42	1.5	1.5	
313	60	3.3	3.3	2.80	.44	0.6	0.7	
501	60	3.3	3.2	2.77	.43	1.4	1.3	
314	59	3.3	3.2	2.61	.43	0.5	0.5	
502	59	3.3	3.2	2.59	.43	0.5	0.5	
212	57	3.2	3.2	2.54	.42	1.1	1.1	
514	57	3.2	3.2	2.53	.43	2.0	2.0*	
509	60	3.3	3.2	2.46	.43	1.0	1.0	
303	56	3.1	3.1	2.39	.42	1.7	1.7*	
205	54	3.0	3.1	2.34	.41	0.9	1.0	
102	53	2.9	3.1	2.31	.41	0.3	0.3	
507	59	3.3	3.1	2.27	.43	0.7	0.7	
302	55	3.1	3.1	2.21	.42	0.8	0.8	
210	55	3.1	3.1	2.18	.42	1.0	1.0	
510	51	2.8	3.1	2.18	.41	0.3	0.3	
108	58	3.2	3.1	2.17	.43	1.4	1.4	
103	52	2.9	3.1	2.14	.41	0.6	0.6	
114	55	3.1	3.0	1.98	.42	1.0	1.0	
508	57	3.2	3.0	1.91	.43	0.5	0.5	
403	51	2.8	3.0	1.80	.40	1.2	1.1	
309	52	2.9	2.9	1.77	.41	0.7	0.7	
** 17 CANDIDATES NOT SHOWN								
306	47	2.6	2.9	1.70	.38	0.8	0.9	
410	54	3.0	2.9	1.59	.42	0.5	0.5	
201	52	2.9	2.9	1.54	.41	0.4	0.3	
315	53	2.9	2.9	1.54	.41	1.4	1.4	
206	49	2.7	2.9	1.51	.39	1.3	1.3	
203	51	2.8	2.8	1.36	.41	0.5	0.5	
209	51	2.8	2.8	1.33	.40	0.9	1.0	
109	53	2.9	2.8	1.29	.42	0.6	0.6	
111	50	2.8	2.8	1.25	.40	1.9	1.9*	
407	49	2.7	2.7	1.04	.39	0.8	0.9	
106	47	2.6	2.7	1.00	.38	0.6	0.6	
402	45	2.5	2.6	.90	.37	1.2	1.3	
207	48	2.7	2.6	.89	.38	1.1	1.2	
307	46	2.6	2.6	.84	.37	1.6	1.6*	
503	48	2.7	2.6	.71	.39	1.0	0.9	
115	46	2.6	2.5	.56	.37	2.5	2.5*	
411	47	2.6	2.5	.48	.38	1.3	1.4	
404	41	2.3	2.5	.46	.34	1.3	1.3	
113	45	2.5	2.5	.42	.37	0.5	0.5	
414	44	2.4	2.5	.41	.36	0.5	0.5	
105	42	2.3	2.4	.36	.35	0.6	0.8	
110	45	2.5	2.4	.34	.37	0.5	0.5	
FAIR AVERAGE OF 2 + 1.65 (SEM) 90% CONFIDENCE								FAIL = 14 CANDIDATES (19%)
204	39	2.2	2.4	.20	.33	0.4	0.4	INSURES THAT NO CANDIDATE WHO SHOULD FAIL WOULD PASS
406	39	2.2	2.4	.20	.33	1.0	1.0	
511	35	1.9	2.3	.16	.31	1.5	1.4	
409	42	2.3	2.3	.09	.35	2.1	2.1*	
413	41	2.3	2.3	.07	.34	0.6	0.7	
107	45	2.5	2.3	.05	.37	0.7	0.7	
215	40	2.2	2.3	.03	.34	1.2	1.2	
202	41	2.3	2.3	-.04	.34	0.5	0.5	
415	39	2.2	2.2	-.17	.33	1.4	1.3	
401	34	1.9	2.1	-.32	.31	1.0	1.1	
FAIR AVERAGE OF 2 TO PASS THE EXAMINATION: SATISFACTORY								FAIL = 3 CANDIDATES (4%)
312	40	2.2	1.9	-.71	.34	0.4	0.4	
213	31	1.7	1.8	-.86	.30	0.7	0.7	
301	26	1.4	1.5	-1.30	.29	1.0	1.0	
Mean	51.9	2.9	2.9	1.79	.40	1.0	1.0	
S.D	9.1	0.5	0.5	1.40	.05	0.5	0.5	

RMSE .41 Adj S.D. 1.34 Separation 3.29 Reliability .92
Fixed (all same) chi-square: 941.3 d.f.: 73 Significance .00

Confidence in Standard	
	Confidence Level
1.0 X(SEM) =	70%
1.3 X(SEM) =	90%
1.65 X(SEM) =	95%
2.00 X(SEM) =	99%

RATING SCALE: 4 = EXCELLENT; 3 = GOOD; 2 = SATISFACTORY
1 = MARGINAL; 0 = UNSATISFACTORY

ALL CANDIDATES RECEIVED 10 RATINGS
*INDICATES RATERS DISAGREED ON CANDIDATE PERFORMANCE

BEST COPY AVAILABLE

References

Cizek, G.J. (1996). Standard Setting Guidelines. Educational Measurement Issues and Practice, 15, 1, 13-21.

Linacre, J.M. (1996). FACETS, a computer program for analysis of examinations with multiple facets. Chicago, IL: MESA Press.

Linacre, J.M. (1989). Many-Facet Rasch Measurement. Chicago, IL: MESA Press.

Lunz, M.E. and O'Neill, T. (1997). A Longitudinal Study of Judge Leniency and Consistency. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Lunz, M.E. and Stahl, J.A. (1993). Impact of Examiners on Candidate Scores: An Introduction to the Use of Multifacet Rasch Model Analysis for Oral Examinations. Teaching and Learning in Medicine, 5, 3, 174-181.

Lunz, M.E., Stahl, J.A. and Wright, B.D. (1994). Interjudge Reliability and Decision Reproducibility. Educational and Psychological Measurement, 54, 4, 913-925.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14, 4, 5-8.

Rasch, G. (1969/1980). Probability models for some intelligence and achievement tests. Chicago, IL: University of Chicago Press.

Wright, B., and Stone, M. (1979). Best Test Design. MESA Press.